FRONTIERS IN GENETICS

HYPOTHESIS AND THEORY ARTICLE

published: 02 April 2014

doi: 10.3389/fgene.2014.00068

A cautionary note on ignoring polygenic background when mapping quantitative trait loci via recombinant congenic strains

J Concepción Loredo-Osti*

Department of Mathematics and Statistics, Memorial University, St. John's, NL, Canada

Edited by:
José M. Âlvarez-Castro, Universidade de Santiago de Compostela, Spain

Reviewed by:
Hongying Dai, Children's Mercy Hospital, USA
Carl Nettelblad, Uppsala University, Sweden

***Correspondence:**
J Concepción Loredo-Osti, Department of Mathematics and Statistics, Memorial University, Henrietta Harvey Building, St. John's, NL A1C 5S7, Canada
e-mail: jcloredoosti@mun.ca

In gene mapping, it is common to test for association between the phenotype and the genotype at a large number of loci, i.e., the same response variable is used repeatedly to test a large number of non-independent and non-nested hypotheses. In many of these genetic problems, the underlying model is a mixed model consistent of one or very few major genes concurrently with a genetic background effect, usually thought as of polygenic nature and, consequently, modeled through a random effects term with a well-defined covariance structure dependent upon the kinship between individuals. Either because the interest lies only on the major genes or to simplify the analysis, it is habitual to drop the random effects term and use a simple linear regression model, sometimes complemented with testing via resampling as an attempt to minimize the consequences of this practice. Here, it is shown that dropping the random effects term has not only extreme negative effects on the control of the type I error rate, but it is also unlikely to be fixed by resampling because, whenever the mixed model is correct, this practice does not allow to meet some basic requirements of resampling in a gene mapping context. Furthermore, simulations show that the type I error rates when the random term is ignored can be unacceptably high. As an alternative, this paper introduces a new bootstrap procedure to handle the specific case of mapping by using recombinant congenic strains under a linear mixed model. A simulation study showed that the type I error rates of the proposed procedure are very close to the nominal ones, although they tend to be slightly inflated for larger values of the random effects variance. Overall, this paper illustrates the extent of the adverse consequences of ignoring random effects term due to polygenic factors while testing for genetic linkage and warns us of potential modeling issues whenever simple linear regression for a major gene yields multiple significant linkage peaks.

Keywords: misspecified genetic models, bootstrapping mixed models, recombinant congenic strains, ignoring random effects, mapping quantitative trait loci

1. INTRODUCTION

For more than four decades, linear mixed models have been used in a wide range of applications because of their conceptual simplicity and flexibility to accommodate correlated sources of variation as well as fixed regressors. A generic linear mixed model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e} \tag{1}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are known incidence matrices, $\boldsymbol{\beta}$ is a vector of unknown fixed regression coefficients, $\boldsymbol{\gamma}$ is a vector of random effects, and $\mathbf{e}$ is the vector of errors. It is also common to assume that $\boldsymbol{\gamma}$ and $\mathbf{e}$ are independent and both have null expectation and finite variances. In many situations, either intentionally or unintentionally, the statistical analysis is carried out ignoring the term $\mathbf{Z}\boldsymbol{\gamma}$ in the model. This practice, although recognized as inefficient, has been thought to be harmless whenever the interest resides solely on a subset of the regression coefficients with the remaining parameters of the model deemed as nuisance. This thought seems to be mostly based on the fact that $\boldsymbol{\beta}^o = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is still an unbiased and consistent estimator of $\boldsymbol{\beta}$. However, it is well known that ignoring $\mathbf{Z}\boldsymbol{\gamma}$ and using ordinary least squares, results in an estimator of $\mathrm{Var}(\boldsymbol{\beta}^o)$ that is biased and inconsistent as well as non-independent of $\boldsymbol{\beta}^o$ (Dhymes, 1978). Of course, this will affect the distribution properties associated with $\boldsymbol{\beta}^o$ under normality or, otherwise, the asymptotic properties of its distribution. It has been suggested that this problem can be mitigated if testing is done through resampling. However, the adverse consequences of dropping the random term from the mixed model is unlikely to be fixed by the use of resampling methods. In this paper, a specific application to genetic mapping via recombinant congenic strains (RCS) of experimental animals is used to illustrate this. Briefly speaking, genetic mapping can be seen as a problem in which the association of one dependent variable (the phenotype) with a large number of potential explicative variables (the marker genotypes) is tested one-by-one or by taking a very small number of markers at once. An RCS panel is a replicable mapping population for which animals within the same strain are considered to be genetically identical and related to different degrees

with animals from other strains. Such an inter-strain relationship results in what is known as the genetic background effect and, whenever this effect is understood as the result of the addition of many components of minuscule effect, the inclusion of a random effects term in the model would be the natural way to account for it.

A mouse panel of RCS is obtained by mating mice from two genetically distinct inbred strains (a donor strain and a recipient strain) followed by two or more rounds of backcrossing to the recipient strain and subsequent sister × brother mating without selection for particular markers or phenotypes for a minimum of 20 generations. The genetic resolution of the panel is controlled by the number of backcrossing rounds. Because of this construction, each strain of an RCS panel can be thought of as an inbred strain in which segments of random length from the genome of a recipient strain have been replaced with the corresponding segments from a donor strain. The main consequence of this breeding scheme is that non-linked genes controlling the same trait are separated and fixed in haplotypes of different strains, allowing the possibility of studying them individually. The standard RCS panel uses two backcross generations and, consequently, the total length of the segments from recipient strain constitute on average the 87.5% of the genome of each strain; the remaining 12.5% represents the total expected length of the replaced genome segments. Without loss of generality, this is the type of RCS considered in this paper. For a more comprehensive description of the RCS and their use in gene mapping see Démant and Hart (1986), Moen et al. (1992), and Fortin et al. (2001b, 2007). Once the RCS panel have been established, the whole panel is genotyped to obtain full characterization of the genome of each strain. Each genotype data set can then be used for the analysis of all individuals of the same strain; this is an important money-saving feature of the design since it does not require of re-genotyping each individual because, except for *de novo* mutations, all pups from the same strain are genetically identical.

Although most mouse geneticists agree that RCS are a powerful resource to map loci associated with complex traits, there is some disagreement on how to do the analysis. Originally, when the use of RCS for genetic mapping was proposed, the core idea was to look into the stain distribution pattern with respect to a phenotype of interest and identify the strain that exhibited the largest deviation from the other strains in the RCS panel and subsequently cross it with the recipient strain to obtain F_1 and F_2 progenies to be analyzed by standard methods (Démant and Hart, 1986; Fortin et al., 2001b). Two examples of the application of this approach are reported in Fortin et al. (2001a) and Müllerová and Hozák (2004). The problem is that contrasting phenotypes from F_1 mice versus the ones from the recipient strain will only be effective for dominant traits, while the power for additive traits will be diminished and lost completely for recessive traits. On the other hand, the analysis of the F_2 mice requires new genotyping, which not only defeats the economic advantages of having developed RCS, but more importantly, because every F_2 individual has different genotype, this approach is not suited for complex quantitative traits when a single measurement may not be reliable enough to determine the phenotype (Moen et al., 1992). Alternatively, there is a designs consisting of taking a sample of mice from each strain and analyzing the whole panel together. Although this approach does not require additional genotyping and has the potential for making more efficient use of the phenotypic variation, also opens more room for analysis pitfalls if the proper model is not used. For example, Joober et al. (2002) uses a QTL mapping procedure equivalent to simple linear regression at the markers ignoring genetic background which, as pointed by Palmer and Airey (2003), it may result in false positive rates far in excess of the nominal value, even when Bonferroni corrections are used. Another common way to address the problem is to use strain averages as the phenotype and treat the panel of means as a backcross dataset for analysis purposes. This is essentially the "interval mapping" procedure proposed by Shao et al. (2010) and equivalent to the one used by Thifault et al. (2008). This approach may substantially reduce the power for RCS panels with reduced number of strains and it does not deal with the fact that the strains, related because their background, may not have the same kinship degree at genomic level and consequently the phenotype means may be not only non-independent but heteroscedastic, as well. Lee et al. (2006) and Camateros et al. (2010) extend the simple linear regression to account for the genetic background by adding a fixed factor ("background proportion" in the first paper; "background indicator" in the second). Although better than ignoring the background, from the genetics standpoint, it is difficult to justify the plausibility of a fixed effects model under the assumption that the background effect is the result of the additive action of many genes of minuscule effect. In fact, I argue that the natural way to model such a background effect consistent with the principles outlined by Fisher (1919) is through the inclusion of a random effects term in the model as implemented in Di Pietrantonio et al. (2010). In this paper, I describe in detail a procedure for the analysis of a quantitative trait locus (QTL) that models the genetic background (assumed to be of polygenic nature) as a random effect term and use this to show how the omission of such a term in the model leads to conclusions that are wrong and inconsistent with the data.

2. MODELS

2.1. THE NAIVE QTL MODEL FOR AN RCS PANEL

In its simplest form, at each marker position m, $m = 1, 2, \ldots, M$, the RCS/QTL model for the ith individual, $i = 1, 2, \ldots, n$, can be written as

$$y_i = \mu + q_{im}\,\xi_m + e_i \tag{2}$$

where y_i denotes the phenotype for the ith individual, ξ_m denotes the major locus effect associated with the mth marker, q_{im} is the indicator of the BB genotype at the mth position which is determined by the RCS data, and the e_is are a set of independent random variables with distribution $\mathcal{N}(0, \sigma^2)$ (AA and BB are the genotypes of the donor and recipient parental strain, respectively). Of course, under an oligogenic model, at most, a handful of ξ_ms should be different from zero. In fact, it is common practice that at the first screening, the estimation is carried out by regression at each marker under the assumption of only one major gene. When the presumption of a dense enough genotyping marker panel is not correct, procedures like modified interval

mapping can be used instead. Variations of the problem include conditioning on a given set of markers. The salient feature of this design is that, at the mth marker position, one looks across the RCS panel and classifies each strain as either AA or BB, since under the model (Equation 2), this is the only source of genetic variation when estimating ξ_m. However, this model ignores the fact that individuals from the same strain are genetically identical (assuming no new mutation at the locus under scrutiny), and strains with the same ancestral background share large portions of their genome so that even without the involvement of a major gene, there is more likely to be reduced variation within strains. In a nutshell, regression mapping works by testing the association of the phenotype with the observed genotype at each marker location so that finding significant linkage at any position implies testing the M null hypotheses, $\xi_m = 0$. Clearly, most of these hypotheses as well as their test statistics are not independent. This may lead to problems in the control of the type I error rate if multiple testing is not addressed properly. Another irregularity results from the fact that with a dense genotyping panel the number of tested hypotheses can by far exceed the sample size. Because of these considerations, p-value estimation by resampling of residuals has been seen as a plausible alternative. For this paper, the problem is addressed through bootstrap.

2.1.1. Computation of p-values

The estimation of genome-wide corrected p-values by resampling requires that under the null hypothesis: (i) each resample is taken from an exchangeable distribution, (ii) the variation of the original sample is preserved through all resamples, and (iii) the genome-wide baseline for the test statistics at each position is the same. The first two requirements are standard for resampling in regression (Davison and Hinkley, 1997; Anderson and Ter Braak, 2003). The last requirement is imposed to ensure that the uncorrected p-values across the genome are comparable (this is particularly important when there are missing genotype data). One way to estimate corrected p-values is to select an ensemble of test statistics whose marginal distribution is the same when the model does not contain any major locus.

Since under model (Equation 2) and the hypothesis of no major gene, the distribution of $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ is exchangeable, resampling from the raw observations will also preserve the variation through the pseudo-observations. This means that in the absence of non-genetic regressors or other non-oligogenic factors, resampling the raw phenotypes either by permutation or through bootstrap will produce similar results. Furthermore, under these premises, basic sampling and hypothesis testing principles indicate that a permutation based procedure will be more efficient and powerful. However, this is not necessarily the case when the premises are removed. Should the model also contain fixed non-genetic regressors, resampling from the leverage-adjusted residuals under the null hypothesis would be a procedure that approximates exchangeability while preserving the original variation of the data. However, under this situation, resampling from leverage-adjusted residuals results in a procedure with acceptable properties only in the bootstrap case (Davison and Hinkley, 1997), while this is not longer guaranteed when resampling via permutation. The main issue is that sampling without replacement magnifies the effects of modest departures from exchangeability. Then, permuting leverage-adjusted residuals may not be good enough (even worst, it may not be valid) and we would require of a much more elaborate and computer intensive procedure to obtain residuals guaranteed to be at least weakly exchangeable so that permutation works properly (see, for example, Kherad-Pajouh and Renaud, 2010). To complete the requirements listed above regarding the possibility of missing genotypes, we propose to use the test statistic defined by the expression

$$z_m = t_m \left(1 - \frac{1}{4\nu_m}\right)\left(1 + \frac{t_m^2}{2\nu_m}\right)^{-\frac{1}{2}} \quad \text{where} \quad t_m = \frac{|\hat{\xi}_m|}{\hat{\sigma}_{\hat{\xi}_m}} \tag{3}$$

and $\hat{\xi}_m$ is the ordinary least squares estimate of ξ_m, $m = 1, 2, \ldots, M$, i.e., z_m is just t_m, our familiar t-statistic with ν_m degrees of freedom, transformed into a z-score (ν_m may vary slightly from marker to marker due to missing data). Another option would be a modified t-statistic t'_m in which the mth estimate of variance s_m^2 used to compute $\hat{\sigma}^2_{\hat{\xi}_m}$ is replaced by s_0^2, the estimate under the null hypothesis. With no missing genotypes the use of any of z_m, t'_m, and t_m would yield approximately the same p-value estimates.

2.1.2. Bootstrap procedure for simple linear regression at the markers

The following bootstrap procedure computes the genome-wide corrected p-values for model (Equation 2) with the test statistic (Equation 3):

STEP 1. At each marker position, m, fit the simple linear regression at the markers model (Equation 2), use (Equation 3) to compute the test statistic z_m, and obtain the genome-wide set of statistics $\mathcal{Z}_M = \{z_m,\ m = 1, 2, \ldots, M\}$. Also, set the genome-wide acceptance count vector to zero.

STEP 2. Sample with replacement from the raw vector of phenotypes, $\mathbf{y} \in \mathbb{R}^n$, to obtain $\mathbf{y}^* \in \mathbb{R}^n$, a bootstrapped full replica of $\mathbf{y}$, and use this vector to compute $z^*_{\max} = \max\{z^*_m\}$, where z^*_m, $m = 1, 2, \ldots, M$, is the test statistic at the mth locus, computed by using $\mathbf{y}^*$, the vector of the pseudo-observations, instead of the original vector of phenotypes.

STEP 3. For each z_m in $\mathcal{Z}_M$, if $z_m \leq z^*_{\max}$, add a unit to the mth entry of the acceptance count vector.

STEP 4. Repeat steps 1 and 2 R times and then compute the estimate of the vector of p-values by dividing the acceptance count vector by R.

This resampling scheme can be seen as an adaptation of a regular regression residuals bootstrapping procedure (Davison and Hinkley, 1997), coupled with Roy's union-intersection principle (Roy, 1953) to control for the genome-wide type I error rate. When applied to the analysis of the RCS panel, this procedure is valid when there is only one observation per strain or when the within-strain variation is negligible. Otherwise, a random term in the model has been neglected and, regardless of $\hat{\xi}_m$ being an

unbiased estimator of ξ_m, the exchangeability requirement cannot be met and the most likely consequence would be an inflated type I error rate. In fact, as per arguments given by Churchill and Doerge (1994) and Churchill and Doerge (2008), this statement is correct not only for the bootstrap and RCS, but also for permutation test procedures applied to any study design involving replicable mapping populations because, as for bootstrap, the Fisher (1935) principle of permutation also relies on exchangeability. For simple experimental designs such as an intercross or a backcross mating, the individual units can safely be assumed to be exchangeable. However, it would be wrong to assume exchangeability for more complicated designs, like advanced intercross, heterogeneous stocks and RCS.

2.2. THE QTL MIXED MODEL FOR AN RCS PANEL

The previous simple linear model (Equation 2) generalizes to a model of the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{q}_m\xi_m + \mathbf{e} \tag{4}$$

where $\mathbf{y}$ represents the phenotype vector, $\mathbf{q}_m$ is a vector with each entry being an indicator variable of the genotype BB at the marker position m with ξ_m being its associated effect (major gene effect), $\boldsymbol{\gamma}$ is a random effects vector associated with the genetic background with $\mathrm{E}(\boldsymbol{\gamma}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\gamma}) = \sigma^2_\gamma\,\boldsymbol{\Delta}_1$, with $\sigma^2_\gamma > 0$ and $\boldsymbol{\Delta}_1$, a positive-definite matrix, both assumed to be constant, although unknown, $\mathbf{X}$ is a matrix of fixed covariates and its corresponding parameter vector $\boldsymbol{\beta}$, $\mathbf{e}$ is a vector of independent and identically distributed random variables representing the error term with $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{e}) = \sigma^2\,\mathbf{I}$. Up to a multiplicative constant, $\boldsymbol{\Delta}_1$ is a function of the length of the segments identical by descent shared amongst strains. For an established RCS panel there are only two possible identity states between pairs of strains at a given locus: either (i) all four alleles are identical by descent ($\boldsymbol{\Delta}_1$ is the matrix holding the pairwise probabilities for this state), or (ii) the strains have different allelic forms and thus identical by descent only amongst themselves. So an estimator of $\boldsymbol{\Delta}_1$ with "a high degree of precision" can be reached. Such an estimator uses only genomic information and does not involve $\mathbf{y}$, so when estimating the parameters, one can assume that $\boldsymbol{\Delta}_1$ is given. Another option is to take the entries of $\boldsymbol{\Delta}_1$ as the expected value of the proportion of the genome shared identical by descent between the respective strains under the RCS panel construction described above, i.e.,

$$\delta_{1ij} = \begin{cases} 1 & \text{if } i = j \\ \frac{15}{16} & \text{if } i \text{ and } j \text{ have the same background} \\ \frac{1}{16} & \text{if } i \text{ and } j \text{ have different backgrounds.} \end{cases} \tag{5}$$

This option, although not the most efficient, does capture the main features of the design and yields a variance structure for the random effects vector that can be exploited in the implementation of the resampling algorithm. For example, if all the strains in the panel under scrutiny have the same background and the simplified expectation-based $\boldsymbol{\Delta}_1$ is used, then the distribution of the vector of random effects is exchangeable. Nonetheless, replacing a genomic-based $\boldsymbol{\Delta}_1$ estimate by its theoretical expectation (Equation 5) implies ignoring important information regarding the correlation of the additive polygenic effects associated to the genetic background.

2.2.1. Estimation

The estimation for the mixed linear model has been extensively discussed in the literature (Harville, 1977; Henderson, 1986). Here we develop an application of these standard methods to the RCS design. Without loss of generality, let us consider the linear mixed model (Equation 1) with $\mathrm{Var}(\boldsymbol{\gamma}) = \sigma^2_\gamma\,\boldsymbol{\Delta}_1$ and $\mathrm{Var}(\mathbf{e}) = \sigma^2\mathbf{I}$. Thus

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \qquad \text{and} \qquad \mathrm{Var}(\mathbf{y}) = \sigma^2\left(\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{I}\right) = \sigma^2\,\boldsymbol{\Sigma}$$

where $\mathbf{G} = \lambda\boldsymbol{\Delta}_1$ and $\lambda = \frac{\sigma^2_\gamma}{\sigma^2}$, i.e., λ represents the signal-to-noise ratio. Under the assumption of no major gene and only polygenic background, λ is related to the heritability coefficient. When $\mathbf{G}$ is known, the best linear unbiased estimator of $\boldsymbol{\beta}$ and the best linear unbiased predictor of $\boldsymbol{\gamma}$ (also known as a shrinkage estimator) can be written as

$$\tilde{\boldsymbol{\beta}} = (\mathbf{W}'\mathbf{W})^{-}\mathbf{W}'\mathbf{v} \quad \text{and} \quad \hat{\boldsymbol{\gamma}} = \mathbf{G}\mathbf{Z}'\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{v} - \mathbf{W}\tilde{\boldsymbol{\beta}}),$$

respectively, where $\mathbf{W} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{X}$ and $\mathbf{v} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{y}$. Also

$$\hat{\sigma}^2 = \frac{1}{N - \mathrm{rank}(\mathbf{W})}(\mathbf{v} - \mathbf{W}\tilde{\boldsymbol{\beta}})'(\mathbf{v} - \mathbf{W}\tilde{\boldsymbol{\beta}})$$

$$\hat{\sigma}^2_\gamma = \frac{1}{\mathrm{rank}(\mathbf{G})}\left(\hat{\boldsymbol{\gamma}}'\mathbf{G}^{-1}\hat{\boldsymbol{\gamma}} + \hat{\sigma}^2\mathrm{tr}(\mathbf{G}^{-1}\mathbf{C})\right)$$

with

$$\mathbf{C} = (\mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{G}^{-1})^{-1} \quad \text{and} \quad \mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Notice that the previous expressions cannot be computed unless the signal-to-noise ratio, λ, is known. A situation of a more practical interest is an iterative procedure on which λ is replaced by its estimate and, once that the estimates of σ^2 and σ^2_γ have been updated, a refinement of the estimate of λ is obtained and so on. This iterative procedure will result in a $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ that are no longer linear, nonetheless, they preserve most of the desirable properties present in their linear counterpart (Jiang, 1998).

2.2.2. Mixed model resampling scheme

Let us now focus our attention toward a resampling scheme appropriate for RCS data under a mixed model. By now, it is obvious that the bootstrap procedure described in the previous section will not work for the mixed model (Equation 4). A crude extension to this procedure would consist of computing

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{Z}\hat{\boldsymbol{\gamma}}$$

and resampling from $\hat{\boldsymbol{\gamma}}$ and $\hat{\mathbf{e}}$ to obtain $\boldsymbol{\gamma}^*$ and $\mathbf{e}^*$ so that the pseudo-observation $\mathbf{y}^*$ could be recovered as

$$\mathbf{y}^* = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{Z}\boldsymbol{\gamma}^* + \mathbf{e}^*.$$

However, it is straightforward to see that these residuals are not exchangeable and they are biased toward zero. Thus, they may not adequately represent the hypothesis tested nor reflect the true variation of the model.

Alternatively, note that when $\boldsymbol{\beta}$ and λ are known, it follows from the model under the null hypothesis that $\mathrm{E}(\mathbf{v}) = \mathbf{W}\boldsymbol{\beta}$ and $\mathrm{Var}(\mathbf{v}) = \sigma^2\,\mathbf{I}$ which implies that the distribution of the vector of residuals, $\boldsymbol{\epsilon} = \mathbf{v} - \mathbf{W}\boldsymbol{\beta}$, is exchangeable. This suggests the following residuals resampling scheme:

(1) given $\tilde{\lambda}$ and $\tilde{\boldsymbol{\beta}}$ obtained under the mixed model without a major gene, i.e., under the null hypothesis, compute $\tilde{\boldsymbol{\Sigma}}$, $\tilde{\mathbf{W}}$ by replacing λ with $\tilde{\lambda}$ and $\boldsymbol{\Delta}_1$ with its genomic-based estimate; then, obtain the leverage-adjusted residuals

$$\tilde{\boldsymbol{\epsilon}} = \mathbf{D}(\tilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathbf{y} - \tilde{\mathbf{W}}\tilde{\boldsymbol{\beta}})$$

where $\mathbf{D}$ is a diagonal matrix with each of the non-zero elements given by $(1 - h_{ii})^{-1}$ and h_{ii} is the ith leverage coefficient;

(2) with replacement, resample from $\tilde{\boldsymbol{\epsilon}} \in \mathbb{R}^n$ to obtain $\boldsymbol{\epsilon}^* \in \mathbb{R}^n$, its bootstrapped replica, and construct the vector of pseudo-observations as

$$\mathbf{v}^* = \tilde{\mathbf{W}}\tilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}^*.$$

If instead of a bootstrap procedure based on leverage-adjusted residuals we want to use a residuals-based permutation procedure, then we need to extend the method of Kherad-Pajouh and Renaud (2010) to get weak exchangeability of residuals. However, when λ is estimated from the data, such an extension is not possible and we would have to rely on approximations. More research is needed to explore this direction.

Outside of a genetics context, there is a number of permutation and bootstrap procedures for mixed models whose objective is testing the components of variance (for example, Fitzmaurice et al., 2007; Sinha, 2009; Lee and Braun, 2012; Samuh et al., 2012). However, they cannot be applied in our case because we are interested in the regression coefficients (or a subset of them) and the variance of the random effects is just nuisance parameter. Incidentally, when testing the components of variance, bootstrap has the edge over most permutation procedures (Samuh et al., 2012).

2.2.3. Bootstrap procedure for the mixed linear model

According to the foregoing argument, generalization to the previous bootstrap procedure to compute the genome-wide corrected p-values for the mixed model (Equation 4) goes as follows:

STEP 0. Compute $\boldsymbol{\Delta}_1$ from the genotype data of the RCS panel, and under the null hypothesis, obtain $\tilde{\lambda}$, $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\Sigma}}$, $\tilde{\mathbf{W}}$ and $\tilde{\boldsymbol{\epsilon}}$ as described in (i) above.

STEP 1. At each marker position, m, fit the model

$$\tilde{\mathbf{v}} = \begin{pmatrix} \tilde{\mathbf{W}} & \tilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathbf{q}_m \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \xi_m \end{pmatrix} + \boldsymbol{\epsilon}. \qquad (6)$$

Of course, this model is equivalent to model (Equation 4), the RCS/QTL mixed model, with λ replaced by $\tilde{\lambda}$. Compute the model parameter estimates with the outlined mixed model procedure as well as the test statistic set $\mathcal{Z} = \{z_m,\ m = 1, 2, \ldots, M\}$ by using Equations (6) and (3); set the acceptance count vector to zero.

STEP 2. Draw a pseudo-observation $\mathbf{v}^*$ by using the proposed resampling scheme in (ii) above and fit the major gene model in model (Equation 6) with $\tilde{\mathbf{v}}$ replaced by $\mathbf{v}^*$ to obtain the set of bootstrapped test statistics $\{z_m^*\}$ and its associated critical value $z_{\max}^* = \max\{z_m^*\}$.

STEP 3. For each z_m in $\mathcal{Z}$, if $z_m \leq z_{\max}^*$, add a unit to the mth entry of the acceptance count vector.

STEP 4. Repeat steps 2 and 3 R times and compute the p-value estimates by dividing the acceptance count vector by R.

To my knowledge, this bootstrap procedure for the analyzing a panel of RCS has not been proposed before Di Pietrantonio et al. (2010) and this paper contains the first detailed derivation and study of its properties. In fact, the resampling methods (mostly conditional permutation) applied to analyze RCS have not used mixed models, but consider the strain effect as fixed which is inconsistent with the hypothesis of a genetic background of polygenic nature or discard information by using only the estimated strain means (for example, Gill and Boyle, 2005; Thifault et al., 2008; Camateros et al., 2010).

3. RESULTS

One straightforward way to show the effect of ignoring the random effects term in a mixed model is by simulation. The idea is to generate a dataset from a model that includes a random term for genetic background and noise, but is free of any major locus. Then compare the p-value profiles (actually, $-\log_{10} p$ profiles) obtained by the use of the naive model (Equation 2) as well as the mixed model (Equation 4). For this simulation study, the genotypes of an RCS panel of 36 strains that were described in Fortin et al. (2001b) were used. The panel originally had 37 lines and 625 microsatellite markers; since then, one line has died out and six markers were removed for reliability reasons. Although a much larger set of single nucleotide polymorphism markers for this RCS panel is also available, I think that this set of 619 markers is enough to show the harmful effects of fitting the wrong model on the inference. Of course, more markers will only exacerbate the problem. For this simulation experiment, six different values for the signal-to-noise ratio parameter λ were chosen (0, $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, 1, and 2). Under a standard additive polygenic model, i.e., a model without major genes, the signal-to-noise parameter is a function of the heritability coefficient (the chosen values correspond to the heritability proportions of 0, $\frac{1}{9}$, $\frac{1}{5}$, $\frac{1}{3}$, $\frac{1}{2}$, and $\frac{2}{3}$, respectively). In every simulation run, a sample of seven individuals from each strain was simulated under the assumption of no major gene, i.e., under model (Equation 4) with $\xi_m = 0$ for all markers, $m = 1, 2, \ldots, M$. The value of σ^2 was fixed for all simulations to 1.175, while $\mathbf{X}\boldsymbol{\beta}$ was fixed as a vector with 7 in all its entries. Simulations for each value of λ were run 1000 times and both methodologies, the mixed model as well as the bootstrapped naive regression at the markers were applied to the simulated datasets with 10, 000 as the number of resamples for every dataset.

In gene mapping studies, a significant peak is defined as the most extreme point of a region beyond the p-value threshold according to some pre-specified genome-wide type I error rate (Churchill and Doerge, 1994). For this study, we use a value of 0.01 or equivalently, a threshold value of 2 on a $-\log_{10} p$-scale. **Tables 1–3** summarize the results of these simulations. As expected, whenever there is not a polygenic term in the model (i.e., $\lambda = 0$), both methodologies produce identical results. However, the picture changes when $\lambda > 0$. In this case, it is quite obvious that ignoring the random effects term has pernicious consequences even for modest levels of λ, the signal-to-noise ratio, while the proposed mixed model method keeps the genome-wide type I error rate relatively close to the nominal value. However, the empirical type I error rates obtained by the proposed procedure seem to increase slightly with λ (**Table 3**). This phenomenon may be due to the fact that the makers used for mapping purposes are also used to estimate the probability of identity by descent between strains and, to a lesser extent, the fact that the the bootstrap procedure is based on residuals computed with λ and $\boldsymbol{\beta}$ estimated from the same data. Nonetheless, the moral of this exercise is that whenever simple regression of a major gene model produces many significant peaks, a warning flag about the model validity should be raised.

Table 1 | Percentage of declared significant peaks with a bootstrap genome-wide adjusted significance level of 0.01 when the proposed mixed model methodology is used.

%		Signal-to-noise ratio (λ)					
		0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	**1**	**2**
Number of significant peaks	0	99.2	98.7	98.9	98.3	98.5	98.4
	1	0.8	0.5	0.5	0.7	0.7	0.4
	2	0	0.2	0.1	0.2	0.5	0.3
	3	0	0.1	0	0.4	0.1	0.3
	4	0	0	0.3	0.1	0.1	0.1
	5	0	0.1	0.1	0.1	0	0.1
	6+	0	0.2	0.1	0.3	0.1	0.4

Estimates based on 1000 simulated datasets for each λ.

Table 2 | Percentage of declared significant peaks with a bootstrap genome-wide adjusted significance level of 0.01 when a naive regression at the markers is used.

%		Signal-to-noise ratio (λ)					
		0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	**1**	**2**
Number of significant peaks	0	99.2	61.1	47.3	38.8	23.9	17.2
	1	0.8	3.9	5.1	5.2	8.1	7.1
	2	0	3.7	4.1	5.1	4.9	6.0
	3	0	1.5	3.5	3.1	3.0	5.3
	4	0	2.5	4.6	3.3	2.1	4.1
	5	0	2.1	5.3	3.2	2.9	2.4
	6+	0	25.2	30.1	41.3	55.1	57.9

Estimates based on 1000 simulated datasets for each λ.

The histogram of a typical dataset obtained by simulation from a model with polygenic effects only would look like the one shown in **Figure 1**. Nonetheless, for this histogram I chose a dataset for which simple linear regression produces a very large number of significant peaks. If a major locus were at play, one would expect to have a well-defined bimodal distribution, so this histogram seems consistent with the generating model of no major gene. However, when we look into the p-value profiles obtained through the model that ignores the genetic background term, instead of profiles consistent with the model we will have something extreme as shown by dashed lines in **Figure 2**. According to the profiles on this figure, one might conclude that all chromosomes have at least one significant peak, fact that does not appear to be supported by the histogram of the data, and more conclusively, this is in conflict with the generating model. If anything, it can be argued that the data distribution may seem a bit skewed, but one may expect that estimation of p-values via bootstrapping of residuals should not be too sensitive to this. Of course, as for bi-modality, skewness may also be caused by a mixture of distributions. However, a very strong peak, as any of the ones spotted on every chromosome, is difficult to conceive without a conspicuous bimodal distribution. Even with the use of robust regression estimates instead of the obtained by regular least squares to minimize the potential impact of outliers on the estimation, these profiles change very little (data not shown). When the missing random effects term is introduced into the model (solid blue

Table 3 | Empirical genome-wide type I error rates obtained via bootstrap in the simulation study (0.01 is the nominal value and the number of simulated datasets for each λ is 1000).

	Signal-to-noise ratio (λ)					
	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	**1**	**2**
Naive regression	0.008	0.389	0.527	0.612	0.761	0.808
Mixed model	0.008	0.013	0.011	0.017	0.015	0.016

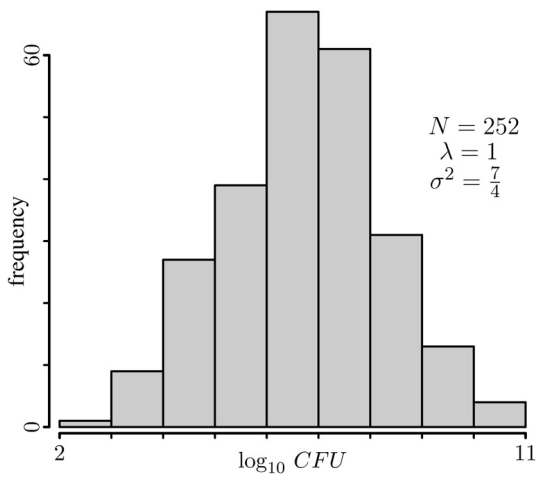

FIGURE 1 | Typical histogram of simulated data. The p-value profiles of the data on this histogram were computed and plotted in **Figure 2**.

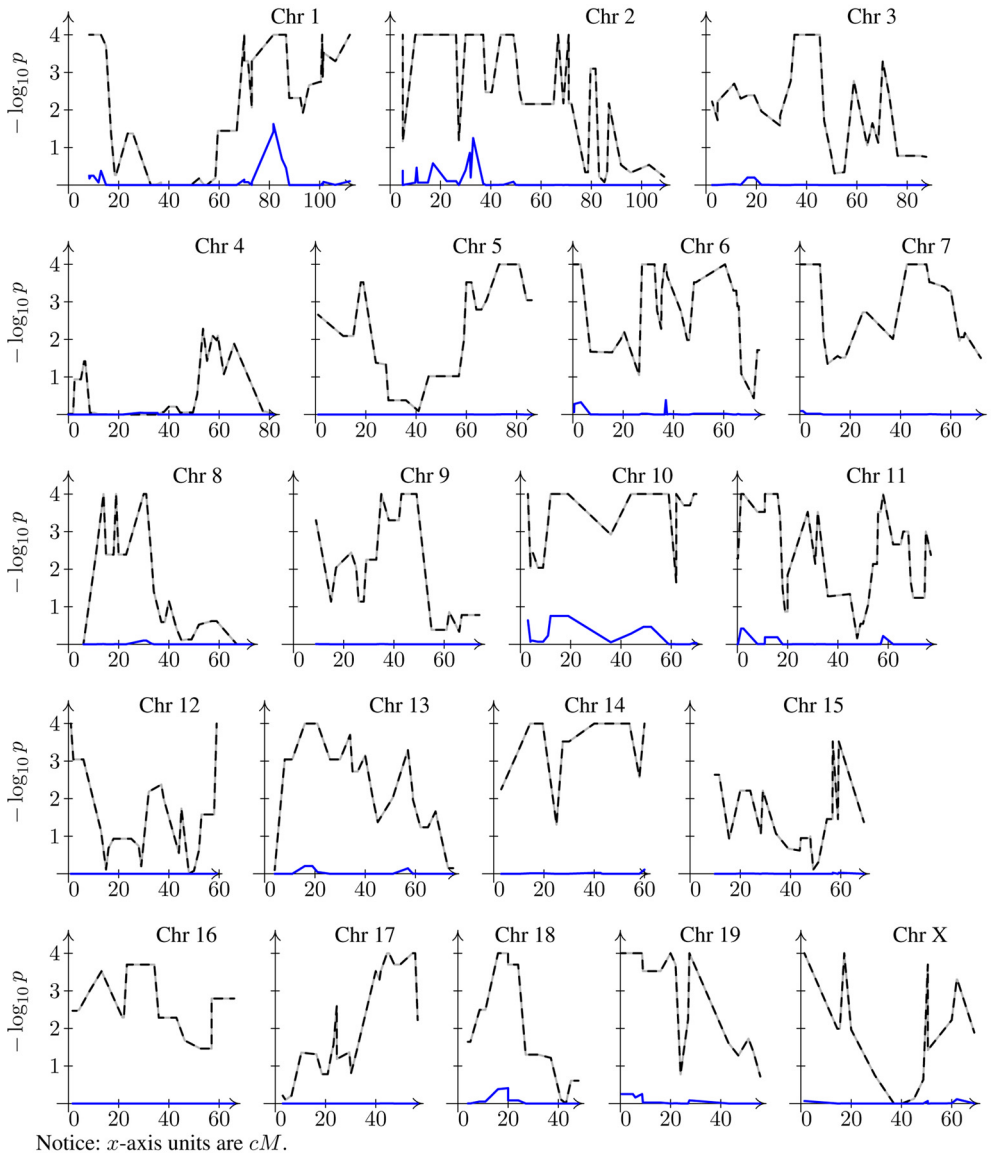

FIGURE 2 | Bootstrap genome-wide corrected *p*-value profiles. Dashed line for naive model (Equation 2) and solid line (sometimes hardly distinguishable from the *x*-axis line) for the mixed model (Equation 6). Note that both profiles have been corrected for multiple testing.

lines in **Figure 2**), *p*-value profiles become consistent with the generating model. Repetition of this exercise on any other simulated datasets yields similar results, although the specific resulting profiles most likely are not be the same.

4. DISCUSSION

This paper proposes a bootstrapping procedure to estimate the *p*-values under a mixed model applied to gene mapping when RCS are used. The method can be easily adapted for other replicable mapping population/designs. This procedure is a generalization of the linear regression bootstrap of residuals coupled with the union-intersection principle aimed to control the genome-wide type I error rate. A simulation study with different values of the signal-to-noise ratio unequivocally shows that when a panel of RCS is used for mapping, ignoring one random effects term in a mixed linear model can have pernicious consequences, resulting in inflated type I error rates and leading to the declaration of significant linkage peaks were no such peaks should be found. The simulation study also shows that the proposed bootstrap procedure seems to produce slightly inflated type I error rates as the signal-to-noise ratio increases. This problem is likely due to the fact that the markers used for mapping are also used to estimate the length of the segments shared identical by descent but also it can be associated with a stronger departure from exchangeability as the ratio increases. In any case, the problem deserves further scrutiny. The proposed bootstrap procedure for mixed models is quite general and can easily be adapted to non-genetic problems.

FUNDING

This work has been supported by the Canadian Institutes of Health Research.

ACKNOWLEDGMENTS

The author expresses gratitude to M. Fujiwara, E. Schurr, T. di Pietrantonio, and K. Morgan for the discussion and comments that substantially improved the manuscript.

REFERENCES

Anderson, M. J., and Ter Braak, C. J. F. (2003). Permutation tests for multi-factorial analysis of variance. *J. Statist. Comput. Simul.* 73, 85–113. doi: 10.1080/00949650215733

Camateros, P., Marino, R., Fortin, A., Martin, J. G., Skamene, E., Sladek, R., et al. (2010). Identification of novel chromosomal regions associated with airway hyperresponsiveness in recombinant congenic strains of mice. *Mamm. Genome* 21, 28–38. doi: 10.1007/s00335-009-9236-z

Churchill, G. A., and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138, 963–971.

Churchill, G. A., and Doerge, R. W. (2008). Naive application of permutation testing leads to inflated type I error rates. *Genetics* 178, 609–610. doi: 10.1534/genetics.107.074609

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511802843

Démant, P., and Hart, A. A. M. (1986). Recombinant congenic strains–a new tool for analyzing genetic traits determined by more than a gene. *Immunogenetics* 24, 416–422. doi: 10.1007/BF00377961

Dhymes, P. J. (1978). *Introductory Econometrics*. New York, NY: Springer. doi: 10.1007/978-1-4612-6292-3

Di Pietrantonio, T., Hernandez, C., Girard, M., Verville, A., Orlova, M., Belley, A., et al. (2010). Strain-specific differences in the genetic control of two closely related mycobacteria. *PLoS Pathog.* 6:e1001169. doi: 10.1371/journal.ppat.1001169

Fisher, R. A. (1919). The correlation between relatives on the supposition of mendelian inheritance. *Trans. R. Soc. Edinb.* 52, 399–433. doi: 10.1017/S0080456800012163

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Fitzmaurice, G. M., Lipsitz, S. R., and Ibrahim, J. G. (2007). A note on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics* 63, 942–946. doi: 10.1111/j.1541-0420.2007.00775.x

Fortin, A., Cardon, L. R., Tam, M., Skamene, E., Stevenson, M. M., and Gros, P. (2001a). Identification of a new malaria susceptibility locus (Char4) in recombinant congenic strains of mice. *Proc. Natl. Acad. Sci. U.S.A.* 98, 10793–10798. doi: 10.1073/pnas.191288998

Fortin, A., Diez, E., Henderson, J. E., Mogil, J. S., Gros, P., and Skamene, E. (2007). "Decoding the genomic control of immune reactions: novartis foundation symposium 281," in *Decoding the Genomic Control of Immune Reactions: Novartis Foundation Symposium 281*, eds G. Bock and J. Goode (Chichester: John Wiley), 141–155. doi: 10.1002/9780470062128

Fortin, A., Diez, E., Rochefort, D., Laroche, L., Malo, D., Rouleau, G. A., et al. (2001b). Recombinant congenic strains derived from A/J and C57BL/6J: a tool for genetic dissection of complex traits. *Genomics* 74, 21–35. doi: 10.1006/geno.2001.6528

Gill, K. J., and Boyle, A. E. (2005). Quantitative trait loci for novelty/stress-induced locomotor activation in recombinant inbred (ri) and recombinant congenic (rc) strains of mice. *Behav. Brain Res.* 161, 113–124. doi: 10.1016/j.bbr.2005.01.013

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* 72, 320–338. doi: 10.1080/01621459.1977.10480998

Henderson, C. R. (1986). Recent developments in variance and covariance estimations. *J. Anim. Sci.* 63, 208–216.

Jiang, J. (1998). Asymtotic properties of the empirical BLUP and BLUE in mixed linear models. *Stat. Sinica* 8, 861–885.

Joober, R., Zarate, J.-M., Rouleau, G.-A., Skamene, E., and Boksa, P. (2002). Provisional mapping of quantitative trait loci modulating the acoustic startle response and prepulse inhibition of acoustic startle. *Neuropsychopharmacology* 27, 765–781. doi: 10.1016/S0893-133X(02)00333-0

Kherad-Pajouh, S., and Renaud, O. (2010). An exact permutation method for testing any effect in balancedand unbalanced fixed effect ANOVA. *Comput. Stat. Data Anal.* 5, 1881–1893. doi: 10.1016/j.csda.2010.02.015

Lee, O. E., and Braun, T. M. (2012). Permutation tests for random effects in linear mixed models. *Biometrics* 68, 486–493. doi: 10.1111/j.1541-0420.2011.01675.x

Lee, P. D., Ge, B., Greenwood, C. M., Sinnett, D., Fortin, Y., Brunet, S., et al. (2006). Mapping cis-acting regulatory variation in recombinant congenic strains. *Physiol. Genomics* 25, 294–302. doi: 10.1152/physiolgenomics.00168.2005

Moen, C. J., Groot, P. C., Dietrich, W., Stoye, J. P., Lander, E. S., and Démant, P. (1992). The recombinat congenic strains for analysis of multigenic traits: genetic composition. *FASEB J.* 6, 2806–2835.

Müllerová, J., and Hozák, P. (2004). Use of recombinant congenic strains in mapping disease-modifying genes. *News Physiol. Sci.* 19, 105–109. doi: 10.1152/nips.01512.2003

Palmer, A. A., and Airey, D. C. (2003). Inappropriate choice of the experimental unit leads to a dramatic overestimation of the significance of quantitative trait loci for prepulse inhibition and startle response in recombinant congenic mice. *Neuropsychopharmacology* 28, 818. doi: 10.1038/sj.npp.1300064

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Stat.* 24, 220–238. doi: 10.1214/aoms/1177729029

Samuh, M. H., Grilli, L., Rampichini, C., Salmaso, L., and Lunardon, N. (2012). The use of permutation tests for variance components in linear mixed models. *Commun. Stat. Theor. Methods* 41, 3020–3029. doi: 10.1080/03610926.2011.587933

Shao, H., Sinasac, D. S., Burrage, L. C., Hodges, C. A., Supelak, P. J., Palmert, M. R., et al. (2010). Analyzing complex traits with congenic strains. *Mamm. Genome* 21, 276–286. doi: 10.1007/s00335-010-9267-5

Sinha, S. K. (2009). Bootstrap tests for variance components in generalized linear mixed models. *Can. J. Stat.* 37, 219–234. doi: 10.1002/cjs.10012

Thifault, S., Sun, Y., Fortin, A., Skamene, E., Lalonde, R., Tremblay, J., et al. (2008). Genetic determinants of emotionality and stress response in AcB/BcA recombinant congenic mice and *in silico* evidence of convergence with cardiovascular candidate genes. *Hum. Mol. Genet.* 17, 331–344. doi: 10.1093/hmg/ddm277

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 October 2013; accepted: 17 March 2014; published online: 02 April 2014.
*Citation: Loredo-Osti JC (2014) A cautionary note on ignoring polygenic background when mapping quantitative trait loci via recombinant congenic strains. Front. Genet. **5**:68. doi: 10.3389/fgene.2014.00068*
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics.